In-Silico Biology:

From Genomics to Organism

SESSION CHAIRS: Robert Berwick and Simon Kasif

DARPA ORGANIZERS: Eric Eisenstadt and Gary Strong

What is the “genomic excitement” about?

- **The entire DNA sequences of many biological organisms are now available on the WEB.**
  For microbial organisms see:
  http://www.tigr.org/tdb/tdb.html
- **Genomic methods are changing biology: thousand of new genes discovered, lateral transfer in bacteria, gene chips, genetic regulation.**
- **The increase in data makes predictive biology effective.**

What is next?

Manuals of Life: Producing integrated databases of biological knowledge.

Bio-interfaces: Developing languages to express knowledge about biological data, processes and experiments.

Virtual Cells: Producing models of biological cells that are consistent with genomic data.

Bio-programming: Creating systems for manufacturing predictable biological systems that are controlled and monitored by a semi-automated in-silico bio-interfaces and virtual cells.

* these research tasks are listed in order of estimated difficulty.

Manuals of Life can take different forms:

- User's guide
- How-to 'built it yourself' kit.
- Repair manual.
- these are not mutually exclusive, all carry their own view of what representations are needed.

Can we build a BioCYC for biology but without making a commitment to a single logical language?

Instead relying on multi-scale, multi-resolution approach that includes probabilistic networks, dynamical systems, logics, simulation languages and other representation mechanisms as appropriate. Similar to engineering approaches for circuit design.

What can we do now:

Engineering Inspired Representations and Algorithms :

- Gene Finding with HMMs
- Protein Function Assignment with Probabilistic Representations (PFAM)
- tRNA scan
- Gene Regulation by Dynamical Systems and Boolean Networks
- Gene Fusion Events by Database Search
- Discovery of co-regulation by Text Search
- Bio-Spice
- MicroArrays
- Experimental Design (e.g, multiplex PCR)
- Fusion of Information Sources (Eisenberg)
- Genetic Switches (Collins)

Gene Finding Observation

- Most genes in public databases were predicted by hidden Markov models, interpolated Markov models, and edit-distance methods. (*).
- Gene recognition technology in microbial DNA is achieving 98% accuracy, a remarkably low error rate for predictive biology systems.
- **(See our system glimmer, http://www.tigr.Org).**

(*) These techniques were developed by the DARPA speech/language understanding initiative.

TALKS IN OUR SESSION

- Whole genome analysis (Salzberg).
  - Detecting foreign DNA: important for detecting potentially dangerous microbial organisms.
  - Genome rearrangement: tuberculosis genome → leprosy genome by inversion + deletions; Important for building and detecting new microbial organisms.
- Predictive biology (Stormo) — a step towards automating the construction of DNA binding proteins.
- Bio-spice – multi-level reasoning about biological systems (Arkin) – a step towards virtual cells.
- A language to describe gene interaction (Brent).

More Talks and Ideas

- Language for describing gene structure and protein structure (Searls).
- Sophisticated analysis of medical literature correlated in a non-trivial manner to results of experiments (Rzhetsky).
- Language for describing protein structure and activity for function prediction (Elber).
- Vertical integration of information sources and data types for constructing on-line bio-manuals.
- Virtual cell.

Virtual Cell

- Virtual cell projects are in progress around the globe.
- Virtual human has been proposed to the academy of sciences by Charles DeLisi.
- Virtual cell construction requires challenging vertical integration across databases, representations and methods.
- DARPA supported the development of the most relevant technologies that include: probabilistic modeling, qualitative reasoning, knowledge mediators, WEB agents, and language modeling.

A short-term methodology:

Computationally well-formed biological questions (protein shape, multiple sequence alignment, motif detection, evolutionary trees, high-performance issues…

How do algorithms scale up or perform in the presence of increasing data – benchmarks, performance analysis…

A high-risk methodology:

Not well formed problems: (virtual cell, automated pathway discovery, gene function, semi-automated genome comparisons, building genetic switches, virtual human…)

How to bring different perspectives to make progress in moving towards effective algorithms – very important to make big advances.

What should DARPA do?

Computational Technique Oriented or Biological Problem Oriented?

- Focus on few computational techniques and representations such as Probabilistic Networks, Logical Rules, Dynamical Systems, Grammars, Language Processing, Approximation Algorithms for Hard Optimization Problems.
- A cut across the most creative ideas to model and possibly control a complex biological system such as the immune system, human pathogens - virtual cells, important biological pathways that cut across species, genetic switches.

Requires a cut across theory and experiments!